Few-shot Single-view 3D Reconstruction with Memory Prior Contrastive Network

Anonymous ECCV submission

Paper ID 192

Abstract. 3D reconstruction of novel categories based on few-shot learning is appealing in real-world applications and attracts increasing research interests. Previous approaches mainly focus on how to design shape prior models for different categories. Their performance on unseen categories is not very competitive. In this paper, we present a Memory Prior Contrastive Network (MPCN) that can store shape prior knowledge in a few-shot learning based 3D reconstruction framework. With the shape memory, a multi-head attention module is proposed to capture different parts of a candidate shape prior and fuse these parts together to guide 3D reconstruction of novel categories. Besides, we introduce a 3D-aware contrastive learning method, which can not only complement the retrieval accuracy of memory network, but also better organize image features for downstream tasks. Compared with previous few-shot 3D reconstruction methods, MPCN can handle the inter-class variability without category annotations. Experimental results on a benchmark synthetic dataset and the Pascal3D+ real-world dataset show that our model outperforms the current state-of-the-art methods significantly.

1 Introduction

Reconstructing 3D shapes from RGB images is valuable in many real-world applications such as autonomous driving, virtual reality, CAD and robotics. Traditional methods for 3D reconstruction such as Structure From Motion (SFM) [29] and Simultaneous Localization and Mapping (SLAM) [2] often require significant efforts in data acquisition. For example, a large number of images need to be captured and the camera parameters need to be calibrated, which limit the applications of these traditional methods.

In recent years, 3D reconstruction from single image based on deep neural networks attracts great research interests. However, previous methods of single-view 3D reconstruction are mostly category-specific [42, 8]. Therefore, they only perform well in the specific training categories. These methods also require a large number of labeled training images, which is time consuming and costly to obtain.

Notably, Tatarchenko et al. [31] shows that single-view 3D reconstruction of specific categories is closely related to image recognition for shape matching. Several simple image retrieval baseline methods can even outperform state-of-the-art 3D reconstruction methods.

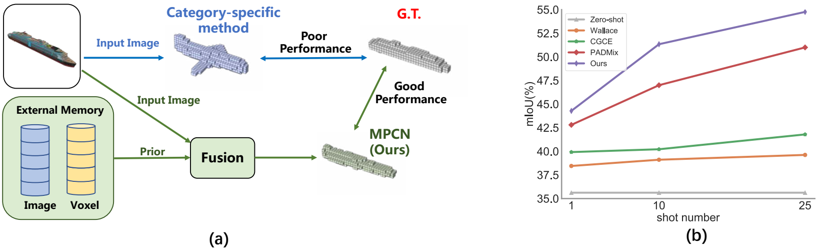

Fig. 1. (a) Novel category 3D reconstruction results on category-specific method and our method combination of prior. (b) The mIoU(%) of current methods against the number of shots. Our MPCN outperforms the SOTA approaches with different shot.

In this paper, we propose a novel category-agnostic model for single-view 3D reconstruction in the few-shot learning settings. In our method, the network parameters are first optimized on some specific object categories. Then given novel categories, the model can be quickly transferred to 3D reconstruction of these categories based on few-shot learning techniques.

To the best of our knowledge, there are mainly three previous works focus on unseen category 3D reconstruction. Wallace et al. [34] present a network for single-view 3D reconstruction with additional shape prior of object categories. However, their shape prior model cannot catch the diversity of shapes within an object category. The authors of Compositional Global Class Embeddings (CGCE) [22] adopt a solution to optimize the shape prior codebooks for the reconstruction of unseen classes. Their model depends on finetuning with additional codebooks, which makes the training process complicated and makes the performance unstable. Pose Adaptive Dual Mixup (PADMix) [6] proposes to apply mixup [46] at the feature level and introduce pose level information, which reaches a new state-of-the-art performance in this task.

In addition, all the works rely on shape prior of specific categories. As a result, additional category annotation is needed to recognize the category of the input image. Then these methods can construct shape prior according to the category annotation, which is not very suitable for category-agnostic 3D reconstruction with novel categories.

The previous works of exploring shape prior for 3D reconstruction of novel categories are insightful and reasonable. However, there still exits a challenge on how to handle shape variety within a novel category in the context of few-shot learning. In this paper, we present a novel deep network model with a memory that can store a shape and its corresponding image as a key-value pair for retrieval. When an image of a novel category is inputted to the network, our deep network can select and combine appropriate shapes retrieved from the memory without category annotation to guide a decoder for 3D reconstruction. In order to adaptively combine the stored shapes as a prior for the downstream 3D reconstruction task, a multi-head attention shape prior module is proposed.

Fig. 1(a) shows the example on novel watercraft category 3D reconstruction performance between traditionally category-specific method [8] and our method.

Besides, we propose a 3D-aware contrastive loss that pulls together the image features of objects with similar shape and pushes away the image features with different 3D shape in the latent feature space, which helps for both organizing the image feature and improving the retrieval accuracy of memory network. Our 3D-aware contrastive loss takes into account the difference of shape as a weighting term to reduce or stress the positiveness of a pair, regardless of the category or instance, as we aim at a category-agnostic 3D reconstruction network.

In summary, our contributions are as follows: We propose a novel Memory Prior Contrastive Network (MPCN) that can retrieve shape prior as an intermediate representation to help the neural network to infer the shape of novel objects without any category annotations.

Our multi-head attention prior module can automatically learn the association between retrieved prior and pay attention to different part of shape prior. It can not only provide prior information between object categories, but also represent the differences within objects in the same category.

The network with both reconstruction loss and a contrastive loss works together for better result. Our improved 3D-aware contrastive loss takes into account of the difference of positive samples, which is more suitable for supervised 3D tasks.

Experimental results on ShapeNet [3] dataset show that our method greatly improves the state-of-the-art methods on the two mainstream evaluation metrics using Intersection over Union and F-score. The reconstruction results on the real-world Pascal3D+ [41] dataset also demonstrate the effectiveness of our method quantitatively and qualitatively.

2 Related Work

2.1 Deep Learning 3D reconstruction

Recently, Convolutional Neural Network (CNN) based single-view 3D reconstruction methods become more and more popular. Using voxels to represent a 3D shape is suitable for 3D CNNs. In the early work 3D Recurrent Reconstruction Neural Network (3D-R2N2) [8], the encoder with a Recurrent Neural Network (RNN) structure is used to predict 3D voxels by a 3D decoder. A follow-up work, 3D-VAE-GAN [39], explores the generation ability of Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) to infer 3D shapes. Marrnet [38] and ShapeHD [40] predict the 2.5D information such as depth, silhouette and surface normal of an input RGB image, and then use these intermediate information to reconstruct the 3D object. OGN[30] and Matryoshka[27] utilize octrees or nested shape layers to represent high resolution 3D volume. Pix2Vox [42] and Pix2Vox++ [43] mainly improve the fusion of multi-view 3D reconstruction. Mem3D[45] introduces external memory for category-specific 3D reconstruction. However, its performance relies on a large number of samples saved during the training process. Its prior module uses a simple RNN, which is unstable to the input sequence. In addition to voxels, 3D shapes can also be represented by point clouds [9, 20, 35], meshes [36, 37] and signed distance fields [44, 21].

2.2 Few-shot Learning

Few-shot learning models can be roughly classified into two categories: metric-based methods and meta-based methods. Metric-based methods mainly utilize Siamese networks [19], match networks [33, 7] or prototype networks [10] to model the distance distribution between samples such that similar samples are closer to each other and heterogeneous samples are far away from each other. Meta-based methods [26][25] and meta-gradient based methods [4, 15, 28] are teaching models by using few unseen samples to quickly update the model parameters in order to achieve generalization.

2.3 Few-shot 3D Reconstruction

Wallace et al.[34] introduce the first method for single-view 3D reconstruction in the few-shot settings. They propose to combine category-specific shape priors with an input image to guide the 3D reconstruction process. However, in their work, a random shape or a calculated average shape is selected for each category as the prior information, which cannot account for shape diversity among objects in a category. In addition, the method does not explicitly learn the inter-class concepts. Compositional Global Class Embeddings (CGCE) [22] adopts a solution to quickly optimize codebooks for 3D reconstruction of novel categories. Before testing on a novel category, the parameters of other modules are fixed. Only the weight vector of codebooks are optimized with a few support samples from the novel category. Therefore, given a new category, CGCE needs to add a new codebook vector for this category and finetune the weight parameters, which makes the whole process complicated and inefficient. Pose Adaptive Dual Mixup(PADMix) [6] proposes a pose adaptive procedure and a three-stage training method with mixup [46]. It reaches a new state-of-the-art but its shape prior module is similar to Wallace [34] and has the drawbacks of complicated three-stage training strategy.

3 Method

Our aim is to design a category-agnostic model, which can achieve superior generalization ability of single-view 3D reconstruction for novel categories with limited support data. Suppose there is a base 3D dataset, defined as $D_b = \{(I_i, V_i)\}_{i=1}^{K}$, where I_i and V_i denote the ith image and its corresponding 3D shape represented using voxels, respectively. K is the number of image and voxel pairs in the dataset. We denote the categories in the base dataset D_b as base categories. Let $D_s = \{(I_i, V_i)\}_{i=1}^{M}$ be another dataset of (image, voxel)

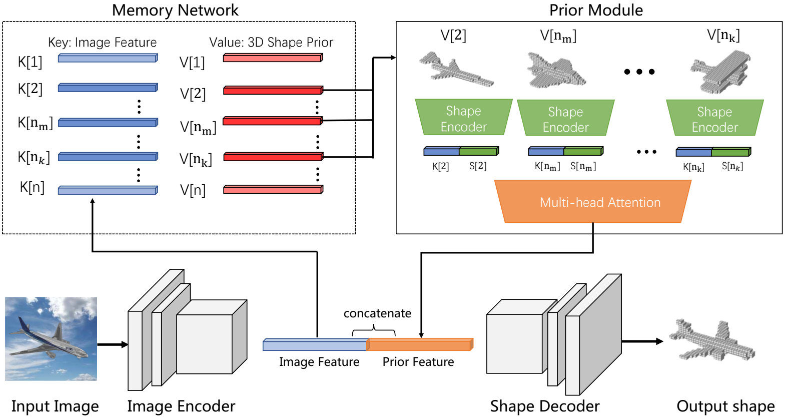

Fig. 2. An overview of the proposed MPCN. In the training stage, we only use base categories to train the model, and insert memory slots into the memory network as alternative prior according to the rules we set. In the test phase, the memory network saves few shot of support set of the novel category to reconstruct 3D volumes of the query set.

pairs with M examples. The categories in D_s are defined as the novel categories, which are different from those in D_b. D_s is called the support set, where $M \ll K$. Meanwhile, there is a large test or query set $D_q = \{(I_i, V_i)\}_{i=1}^{N}$ with N examples. Examples in D_q and D_s are all within the novel categories. Note that we only use D_b and D_s for training. The support set D_s can be used as prior information. We hope that the model can be designed to be category-agnostic and achieve good performance in the query set D_q.

3.1 Memory Network

In most previous works on single-view 3D reconstruction, the shape prior information is learned from the model parameters, which leads to category collapse when transferring to novel categories. As mentioned in [31], such kind of model makes the problem degenerated into a classification task. To alleviate this issue, directly using 3D shapes as priors is an intuitive and effective way. As shown in Figure 2, we adopt an explicit key-value memory network to store and calculate shape priors. In the training and testing stages, the CNN features of the input image is extracted by a 2D encoder. Then a retrieval task is performed, where the keys of the samples stored in the memory network are compared to the query vector and the corresponding Top-k retrieved shapes are sent to the prior module for generating prior features. Specifically, as shown in Eq. (1), the input image I_q is first encoded by a 2D encoder E_{2D}, then the image features and shape prior features are concatenated. Finally, the 3D shape is inferred by a 3D decoder D_{3D}.

$$pr = D_{\text{3D}}(\text{Concatenate}(E_{\text{2D}}(I_q), \text{prior feature})), \qquad (1)$$

where pr is the final predict volume.

Memory Store The external memory module is a database of experiences. Each column in the memory represents one data distribution. In MPCN, two columns of structures in the form of key-value are stored. A key is a deep feature vector of an image, and its value is the corresponding 3D shape represented using voxels. Each memory slot is composed of [image feature, voxel], and the memory module database can be defined as $\mathcal{M} = \{(I_i, V_i)\}_{i=1}^{m}$, where m represents the size of the external memory module. We use a simple but effective memory storage strategy to store data in a limited memory size. During training stage, when generating a target shape with MPCN, we calculate the distance $d(pr, gt)$ between all the samples' prediction and target shape of a batch as in Eq. (2):

$$d(pr, gt) = \frac{1}{{r_v}^3} \sum_{i,j,k} (pr_{(i,j,k)} - gt_{(i,j,k)})^2, \tag{2}$$

where r_v is the resolution of 3D volumes, gt is the ground truth volume. For a sample (I_k, V_k) , if $d(pr, gt)$ is greater than a specified threshold δ, we consider that the current network parameters and the prior have poor reconstruction performance on this shape. So we insert (I_k, V_k) into the external memory module and store it as a memory slot in order to guide the reconstruction of similar shapes in the future. We maintain an external memory module similar to the memory bank(queue). When the memory is full, the memory slot initially added to the queue will be replaced by later one. This makes sense because the later image features are updated with iteration of the model training.

Memory Reader Each row of the external memory database represents a memory slot. The retrieval of the memory module is based on a k-nearest neighbor algorithm. When comparing the CNN features of the current query image and the image features of all slots in the memory network, we use the Euclidean metric to measure the differences. In order to obtain the distance between the query matrix and the key matrix of the memory network conveniently, we use the effective distance matrix computation to calculate the matrix of Euclidean distance as shown in Eq. (3):

$$\text{Distance} = \|Q\| + \|K\| - 2 * QK^T, \tag{3}$$

where $Q \in \mathbb{R}^{b \times 2048}$ is query matrix, $K \in \mathbb{R}^{m \times 2048}$ is memory-key matrix, b is the batch size, and m is the memory size.

After calculating the distance, we select the nearest k retrieval results as prior information, which is defined as $R = \{(I_i, V_i)\}_{i=1}^{k}$. $\{I_i\}_{i=1}^{k}$ represents k retrieved image features, $\{V_i\}_{i=1}^{k}$ represents k retrieved voxels. When the first batch is searched, the memory will be empty. However, we will take out k all-zero tensors, which increases the robustness of the model to some extent. Even without the shape prior as a guide, our model should reconstruct the 3D model according to the 2D features of the image.

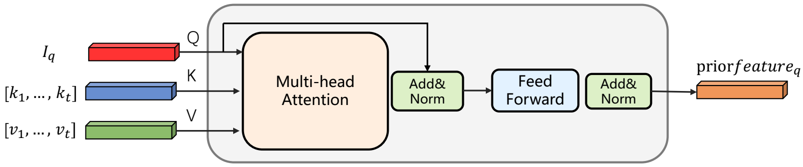

Fig. 3. An overview of the proposed Memory Prior Module.

3.2 Prior Module

The prior module can first obtain the set $\{(I_i, V_i)\}_{i=1}^{k}$ retrieved by the external memory module from the previous step. Note that shape volume is the original voxel saved at this stage, and its size is 32^3. So the model needs to extract shape features by a 3D shape encoder before the downstream processing.

$$k_i = I_i, v_i = \text{Encoder3D}(V_i), \tag{4}$$

$$Q = I_q W_q, K = k_i W_k, V = v_i W_v, \tag{5}$$

$$e_q = Q + \text{LayerNorm}(\text{MHA}(Q, K, V)), \tag{6}$$

$$prior\ feature_q = e_q + \text{LayerNorm}(\text{FFN}(e_q)), \tag{7}$$

In previous works, only 3D voxels are regarded as the prior features. In contrast, as shown in Figure 3, we use the attention based architecture to extract shape prior features by exploring the association between image features and 3D shape. Concretely, we take the query image feature I_q as the query, the retrieved image feature $\{I_i\}_{i=1}^{k}$ as the key, and its corresponding 3D shape feature $\{v_i\}_{i=1}^{k}$ as the value. As shown in Eq. (5), we first use three separate linear layers parameterized by W_q, W_k and W_v to extract query, key, value embedding Q, K and V.

Then the embeddings are forwarded to the multi-head attention(MHA)[32] and layer normalization(LayerNorm) module[1] to perform cross-attention between the query and every key. The output of the attention is fused to the original input query embedding to get enhanced feature e_q. Afterward, the obtained features e_q are sent into feed-forward network(FFN) and layer normalization(LayerNorm). The output $priorfeature_q$ is obtained by adding up the feed-forward module output with residual connection as in Eq. (6) and Eq. (7).

3.3 3D-Aware Contrastive Learning Method

We believe that Memory Prior Network can work effectively mainly based on the accuracy of 2D image embedding retrieval. In order to improve the retrieved prior accuracy provided by memory network, previous work generally used triple loss[13] to optimize encoder, which is effective for simple classification problems [16] [47]. However, for 3D reconstruction task, the triplets need to construct positive and negative samples according to the threshold of specific shape difference, which is an empirical and troublesome step [45].

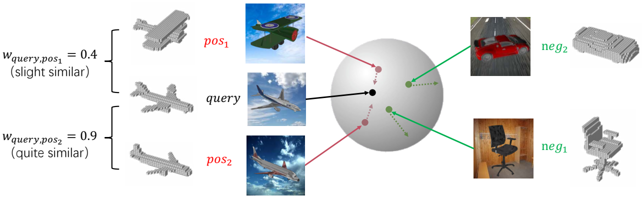

Fig. 4. An example of 3D-Aware Contrastive Loss, which pulls together the positive samples with similar 3D shape(e.g., pos_1 and pos_2) by different *weight*, and pushes apart the negative samples with different 3D shape(e.g., neg_1 and neg_2).

Recently, contrastive learning method [5][11] train an image encoder maximizing the similarity of different transformations of the same sample and minimizing the similarity with other samples. The proposed loss $\mathcal{L}_{infoNCE}$ [11] achieves great success in self-supervised representation learning. However, their success depends on the large batch size and memory occupation, and taking all of samples in the same batch as negative pairs may be wrong for supervised tasks. In addition, supervised contrastive loss[17] tries to solve this problem but it mainly focuses on simple classification problem.

We hope to design a loss that can adaptively pull together the image embedding pairs with similar 3D shape and push away the image embeddings with different 3D shapes. Therefore, as shown in Fig. 4. we introduce an improved 3D-aware contrastive loss, which considers the positivity of positive pairs. Concretely, for each image pair (q, k), we calculate the distance between the associated 3D shape $d(V_q, V_k) \in [0, 1]$, we take (q, k) as a positive pair if $d(V_q, V_k) < \delta$, then a weight is calculated by $w_{q,k} = (1 - d(V_q, V_k) \times \gamma)$, which is considered as the important weight of the positive pair in our 3D-aware contrastive loss:

$$\mathcal{L}_{3DNCE} = -log \frac{\sum_{p \in [1..M]} w_{q,p} \cdot exp(f_q \cdot f_p / \tau)}{M \cdot \sum_{k \in [1..N]} exp(f_q \cdot f_k / \tau).} \tag{8}$$

Where $d(V_q, V_p)$ is the same as Eq (2), q is a query image, p is the positive samples of q according to $d(V_q, V_p) < \delta$, $\sum_{k \in [1..N]}$ mean the samples in the same batch with q, and f is image encoder. Intuitively, the more similar the 3D shape of two objects (q, p) is, the greater its weight $w_{q,p}$ is, and the closer the image features of the two objects are.

3.4 Training Procedure in Few-shot settings

We adopt a two-stage training strategy. In the first stage, we train the initialization model on the base category data D_b. In the second stage, we use few-shot novel category samples in support set D_s to finetune the network. Both stage adopt the training method based on episodes as shown in the Algorithm 1. At the beginning of each epoch, all slots of the memory are cleared to ensure that the new round of training can re-determine which samples are inserted into the

Algorithm 1 Training algorithm

```
for epoch in epochs do
    flush memory slots
    for batch_idx in range(max_episode): do
        Load query images, target shape from train set $D_b$
        embed2d = Encoder2d(query image)
        Key, Value, Distance = Top-K(embed2d)
        embedPrior = Prior(embed2d, Key, Value, Distance)
        embed = concatenate(Embed2d, embedPrior)
        predict = Decoder3d(embed)
        d = computeDis(predict, target shape)
        if $d > \delta$: then
            insert(image, voxel) to external memory
        end if
        Train on predict and target with backprop
    end for
end for
```

memory module according to our memory store strategy. For test phase, we first insert samples in support set D_s to the memory module as candidate prior information according to the few-shot settings. Then it follows the same steps as training stage to predict 3D shape in query set D_q. Finally, the reconstruction results are evaluated by evaluation metric.

3.5 Architecture

Image Encoder The 2D encoder shares the same ResNet backbone [12] as that of CGCE [22]. Then our model follows with three layers of convolution layer, batch normalization layer and Relu layer. The convolution kernel size of the three convolution layers is 3^2 with padding 1. The last two convolution layers are then followed by a max pooling layer. The kernel size of the pooling layer is 3^2 and 2^2, respectively. The output channels of the three convolution layers are 512, 256 and 128, respectively.

Prior Module The 3D shape encoder of the prior module includes four convolution layers and two max-pooling layers. Each layer has a LeakyRelu activation layer, and the convolution kernel sizes are 5^3, 3^3, 3^3 and 3^3, respectively. The output Q,K,V embedding dimension of the Linear layer is 2048. The size of key and value of the attention module are both 2048. This module has 2 heads in attention blocks. Finally, the prior feature dimension is 2048.

Shape Decoder There are five 3D deconvolution layers in this module. The convolution kernel size is 4^3, stripe is 2^3, and padding is 1. The first four convolutions are followed by batch normalization and Relu, and the last is sigmoid function. The output channels of the five convolution layers are 256, 128, 32, 8 and 1, respectively. The final output is a voxel representation with size 32^3.

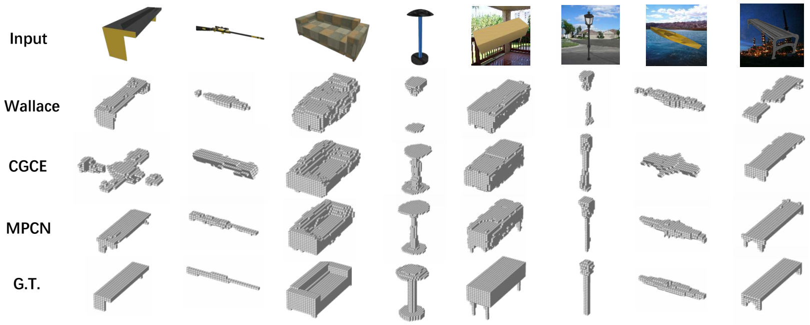

Fig. 5. Examples of single-view 3D Reconstruction on novel category of ShapeNet with shot-10. We show examples with clean background and with random background.

3.6 Loss Function

Reconstruction Loss For the 3D reconstructions network, both the reconstruction prediction and the ground truth are based on voxel. We follow previous works [34, 22, 42, 43] that adopt the binary cross entropy loss as our loss function:

$$\mathcal{L}_{rec} = \frac{1}{r_v^3} \sum_{i=1}^{r_v^3} [gt_i \log(pr_i) + (1 - gt_i) \log(1 - pr_i)], \tag{9}$$

where r_v represents the resolution of the voxel space, pr and gt represent the predict and the ground truth volume.

Total Loss The MPCN is trained end-to-end with the reconstruction loss and 3D-aware contrastive loss together as following:

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \lambda \mathcal{L}_{3DNCE} \tag{10}$$

where λ is hyperparameter, which is set to 0.01 in this work.

4 Experiment

4.1 Experimental setup

Dataset We experiment with the **ShapeNet** dataset [3]. The setting of this dataset is the same as [34]. Seven of 13 categories are designated as the base classes that are used for training: airplanes, cars, chairs, displays, phones, speakers and tables. The other categories are set as novel classes for testing. To fairly compare with the previous works, we use the same dataset split as in [34] and CGCE [22]. The datasets are provided by [8] which are composed with 137×137 RGB images and $32 \times 32 \times 32$ voxelized representations. **Pascal3D+** [41] dataset has 12 different categories. It provides approximate 3D annotations for Pascal

VOC2012 and ImageNet [41]. Each category has about 10 CAD models, which are generally not used for training directly. We finetune our MPCN with 8 categories of Pascal3D+, and test it on four categories: bicycle, motorbike, bottle and train. For fair comparison, we use Binvox [23] tool to render the voxel representation from the CAD model, and the voxel resolution is also 32 × 32 × 32.

Table 1. Comparison of single-view 3D object reconstruction on novel class of ShapeNet at 32^3 resolution with different available shot.We report the **mean IoU** per category. The best number for each category is highlighted in bold.

	0-shot	1-shot				10-shot				25-shot			
category	B0	Wallace	CGCE	PADMix	MPCN	Wallace	CGCE	PADMix	MPCN	Wallace	CGCE	PADMix	MPCN
cabinet	0.69	0.69	0.71	0.67	**0.72**	0.69	**0.71**	0.66	0.68	0.69	0.71	0.68	**0.74**
sofa	0.52	0.54	0.54	0.54	**0.57**	0.54	0.54	0.57	**0.60**	0.54	0.55	0.59	**0.65**
bench	0.37	0.37	0.37	0.37	**0.39**	0.36	0.37	**0.41**	**0.41**	0.36	0.38	0.42	**0.45**
watercraft	0.28	0.33	0.39	**0.41**	**0.41**	0.36	0.41	0.46	**0.54**	0.37	0.43	0.52	**0.55**
lamp	0.19	0.20	0.20	**0.29**	**0.29**	0.19	0.20	0.31	**0.32**	0.19	0.20	0.32	**0.37**
firearm	0.13	0.21	0.23	**0.31**	0.24	0.24	0.23	0.39	**0.52**	0.26	0.28	0.50	**0.52**
mean	0.36	0.38	0.40	0.43	**0.44**	0.40	0.41	0.47	**0.51**	0.41	0.43	0.51	**0.54**

Evaluation Metrics For fair comparison, we follow previous work [34][22][42] using **Intersection over Union (IoU)** as the evaluation metrics. The IoU is defined as following:

$$\mathrm{IoU} = \frac{\sum_{i,j,k} \mathcal{I}(\hat{p}_{(i,j,k)} > t)\mathcal{I}(p_{(i,j,k)})}{\sum_{i,j,k} \mathcal{I}[\mathcal{I}(\hat{p}_{(i,j,k)} > t) + \mathcal{I}(p_{(i,j,k)})]}, \tag{11}$$

where $\hat{p}_{(i,j,k)}$ and $p_{(i,j,k)}$represent the predicted possibility and the value of ground truth at point (i, j, k), respectively. $\mathcal{I}$ is the function that is one when the requirements are satisfied. t represents a threshold of this point, which is set to 0.3 in our experiment.

Implementation details We used 224 × 224 RGB images as input to train the model with a batch size of 16. Our MPCN is implemented in PyTorch [24] and trained by the Adam optimizer [18]. We set the learning rate as $1e-4$. The δ and γ are set to 0.1 and 10. The k of the retrieval samples Top-k is 5. The memory size m is set to 4000 in the training stage, and only 200 in the testing stage.

Baseline We compare our proposed MPCN with three state-of-the-art methods: Wallace[34], CGCE [22] and PADMix[6]. We also follows the zero-shot lower baseline in CGCE [22]. Zero-shot refers to the result of training on the base categories with only single-image and testing directly on the novel class without any prior. Image-Finetune method refers to training on the base categories and finetuning the full network with few available novel categories samples.

4.2 Results on ShapeNet dataset

We compare with the state-of-the-art methods on the ShapeNet novel categories. Table 1 shows the IoUs of our MPCN and other methods. Results in Table 1 from other models are taken from [6]. For few-shot settings, we follow the evaluation in CGCE [22] and PADMix[6] shown the results in the settings of 1-shot, 10-shot and 25-shot. It can be seen that our method is much better than the zero-shot baseline respectively, and greatly outperforms SOTA's methods. Experimental results show that MPCN has great advantages when the shot number increases, mainly because it can retrieve prior information more related to the target shape. Even in the results of 1-shot, there are some improvements, mainly because our model can select the most appropriate prior of shapes as well as image features, and use the differences of other shapes to exclude other impossible shapes. Our MPCN results are shown in Figure 4. It can be seen that our model can obtain satisfactory reconstruction results for novel categories than any other SOTA methods even when the angles of input images are different.

Table 2. Comparison of single-view 3D object reconstruction on Pascal3D+ at 32^3resolution. We report both the mean IoU of every novel category. The best number is highlighted in bold.

	bicycle	motorbike	train	bottle	mean
Zero-shot	0.11	0.27	0.35	0.10	0.2074
Image-Finetune	0.20	0.28	0.35	0.32	0.2943
Wallace [34]	0.21	0.29	**0.40**	0.43	0.3324
CGCE [22]	0.23	0.33	0.37	0.35	0.3223
MPCN(ours)	**0.28**	**0.39**	0.37	**0.46**	**0.3748**

4.3 Results on Real-world dataset

In order to compare with Wallace et al. [34] further, we also conducted experiments on the real-world Pascal3D+ dataset [41]. Firstly, the model is pre-trained on all 13 categories of the ShapeNet dataset [3]. Then the model is finetuned with Pascal3D+ base category dataset, and the final test set is selected from the four novel categories. Note that the experiment is set of 10-shot. The experimental results show that our method outperform zero-shot baseline by 16.74%, also it is the best compared with SOTA methods. Especially for bicycle and motorbike categories with large shape difference and variability, our model perform best. But for the category with subtle shape difference (*e.g.*, train), the reconstruction results tend to align to the average of prior shape, so Wallace [34] shows marginal improvement over ours. But note that our MPCN selects prior information by memory network automatically, while Wallace et al. [34] need the category annotations of the input images and choose shape priors manually. The experimental results are shown in the Table 3.

Table 3. The effectiveness of the different modules and losses. All the results are tested on novel categories of ShapeNet and Pascal3D+ with 5-shot. We report the mean IoU(%) of novel categories. The best number is highlighted in bold.

	MHA	LSTM	Random	Average	Finetune	$\mathcal{L}_{3DNCE}$	$\mathcal{L}_{infoNCE}$	ShapeNet	Pascal3D+
Zero-shot								35.68	20.74
Image-Finetune					✓			40.82	27.49
MPCN-Top1	✓				✓	✓		39.95	31.65
w LSTM		✓				✓		40.52	25.63
w Random prior			✓			✓		35.52	22.05
w average				✓		✓		41.32	29.85
w $\mathcal{L}_{infoNCE}$	✓						✓	42.98	27.98
w/o $\mathcal{L}_{3DNCE}$	✓							43.05	28.76
w/o finetune	✓					✓		45.82	30.32
ours	✓				✓	✓		**47.53**	**33.54**

5 Ablation Study

In this part, we evaluate the effectiveness of proposed modules and the impact of different losses. We choose the setup of 5-shot on both ShapeNet and Pascal3D+ dataset for comparative experiment if not mentioned elsewhere.

Retrieval or Reconstruction In order to prove that our method is superior to the retrieval method, we just take the highest similarity retrieved shape as the target shape. That is the result of MPCN-Top1 in Table 3. In addition, Image-Finetune method is also shown for comparison. The results in Table 3 shows that our MPCN outperforms any upper retrieval or finetune methods in the few-shot settings.

Analysis of Prior Module We analyze the prior extraction module based on attention in MPCN. Because previous methods using external memory network adopt LSTM [14] in the shape prior fusion stage, we replace the attention part of MPCN with LSTM (w LSTM) for the purpose of comparison. We also compare the average-fusion (w average) of the retrieved Top-5 object volume features. In addition, the random initialization of prior vectors (w Random prior) is also compared in the experiment.

The experimental results in Table 3 show that our prior module plays an important role for guiding the reconstruction of 3D objects. The fusion of Top-5 average method and random prior obviously cannot make full use of similar 3D volumes. Our attention module can capture the relevance of different 3D objects better than LSTM, and shows more powerful ability of inferring 3D shapes by using prior information in the few-shot settings. Fig. 6 illustrates some shape priors selected by our model. We demonstrate that the multi-head attention module can adaptively detect the proper parts of the retrieved shapes for 3D reconstruction of novel categories.

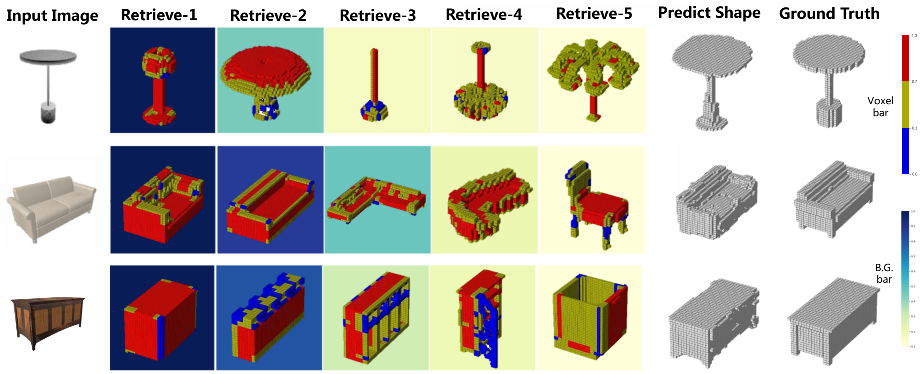

Fig. 6. Visualization of features maps from the retrieved 3D volumes and the corresponding reconstructions.

Analysis of Loss and Finetune In order to further prove the effectiveness of our proposed 3D-aware contrastive loss, we remove $\mathcal{L}_{3DNCE}$ in the training stage, as (MPCN w/o $\mathcal{L}_{3DNCE}$) shown in Table 3. Besides, we replace the improved $\mathcal{L}_{3DNCE}$ with the traditional contrastive loss, as shown (MPCN w $\mathcal{L}_{infoNCE}$) in Table 3. The results show that our comparison $\mathcal{L}_{3DNCE}$ has a great contribution to the improvement of experimental effect, mainly because it not only improves the retrieval accuracy of memory a prior module, but also makes the intent space of 2D representation more reasonable. In addition, using few-shot novel category samples in support set D_s to finetune the network (w finetune) in the second training stage is also important for the performance.

6 Conclusion and Future Works

In this paper, we propose a novel category-agnostic model for 3D object reconstruction. Inspired by the novel 3D object recognizing ability by human-beings, we introduce an external memory network to assist in guiding the object to reconstruct the 3D model in few-shot settings. Compared with the existing methods, our method provides an advanced module to select shape priors, and fuses shape priors and image features to reconstruction 3D shapes. In addition, a 3D-aware contrastive method is proposed for encode 2D latent space, which may be used for other supervised tasks of 3D vision. The experimental results show that our MPCN can outperform existing methods on the ShapeNet dataset and the Pascal3D+ dataset under the settings of few-shot learning.

In future work, we will work on improving the resolution of the reconstructed 3D objects for topologically complex shapes. In addition, we plan to explore the 3D-aware contrastive method on other 3D tasks, such as shape completion or 3D segmentation.

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
2. Cadena, C., Carlone, L., Carrillo, H., Latif, Y., Scaramuzza, D., Neira, J., Reid, I., Leonard, J.J.: Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. IEEE Transactions on robotics **32**(6), 1309–1332 (2016)
3. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015)
4. Chen, R., Chen, T., Hui, X., Wu, H., Li, G., Lin, L.: Knowledge graph transfer network for few-shot recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 10575–10582 (2020)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
6. Cheng, T.Y., Yang, H.R., Trigoni, N., Chen, H.T., Liu, T.L.: Pose adaptive dual mixup for few-shot single-view 3d reconstruction. arXiv preprint arXiv:2112.12484 (2021)
7. Choi, J., Krishnamurthy, J., Kembhavi, A., Farhadi, A.: Structured set matching networks for one-shot part labeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3627–3636 (2018)
8. Choy, C.B., Xu, D., Gwak, J., Chen, K., Savarese, S.: 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In: European conference on computer vision. pp. 628–644. Springer (2016)
9. Fan, H., Su, H., Guibas, L.J.: A point set generation network for 3d object reconstruction from a single image. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 605–613 (2017)
10. Gao, T., Han, X., Liu, Z., Sun, M.: Hybrid attention-based prototypical networks for noisy few-shot relation classification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 6407–6414 (2019)
11. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2020)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
13. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
14. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
15. Jeong, M., Choi, S., Kim, C.: Few-shot open-set recognition by transformation consistency. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12566–12575 (2021)
16. Kaiser, Ł., Nachum, O., Roy, A., Bengio, S.: Learning to remember rare events. arXiv preprint arXiv:1703.03129 (2017)
17. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. arXiv preprint arXiv:2004.11362 (2020)

18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
19. Koch, G., Zemel, R., Salakhutdinov, R., et al.: Siamese neural networks for one-shot image recognition. In: ICML deep learning workshop. vol. 2. Lille (2015)
20. Lin, Y., Wang, Y., Li, Y., Wang, Z., Gao, Y., Khan, L.: Single view point cloud generation via unified 3d prototype. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 2064–2072 (2021)
21. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4460–4470 (2019)
22. Michalkiewicz, M., Parisot, S., Tsogkas, S., Baktashmotlagh, M., Eriksson, A., Belilovsky, E.: Few-shot single-view 3-d object reconstruction with compositional priors. In: European Conference on Computer Vision. pp. 614–630. Springer (2020)
23. Nooruddin, F.S., Turk, G.: Simplification and repair of polygonal models using volumetric techniques. IEEE Transactions on Visualization and Computer Graphics **9**(2), 191–205 (2003)
24. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32**, 8026–8037 (2019)
25. Ramalho, T., Garnelo, M.: Adaptive posterior learning: few-shot learning with a surprise-based memory module. In: International Conference on Learning Representations (2018)
26. Ravichandran, A., Bhotika, R., Soatto, S.: Few-shot learning with embedded class models and shot-free meta training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 331–339 (2019)
27. Richter, S.R., Roth, S.: Matryoshka networks: Predicting 3d geometry via nested shape layers. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1936–1944 (2018)
28. Satorras, V.G., Estrach, J.B.: Few-shot learning with graph neural networks. In: International Conference on Learning Representations (2018)
29. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016)
30. Tatarchenko, M., Dosovitskiy, A., Brox, T.: Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2088–2096 (2017)
31. Tatarchenko, M., Richter, S.R., Ranftl, R., Li, Z., Koltun, V., Brox, T.: What do single-view 3d reconstruction networks learn? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3405–3414 (2019)
32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
33. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. Advances in neural information processing systems **29**, 3630–3638 (2016)
34. Wallace, B., Hariharan, B.: Few-shot generalization for single-image 3d reconstruction via priors. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3818–3827 (2019)

35. Wang, J., Sun, B., Lu, Y.: Mvpnet: Multi-view point regression networks for 3d object reconstruction from a single image. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8949–8956 (2019)
36. Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W., Jiang, Y.G.: Pixel2mesh: Generating 3d mesh models from single rgb images. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 52–67 (2018)
37. Wen, C., Zhang, Y., Li, Z., Fu, Y.: Pixel2mesh++: Multi-view 3d mesh generation via deformation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1042–1051 (2019)
38. Wu, J., Wang, Y., Xue, T., Sun, X., Freeman, B., Tenenbaum, J.: Marrnet: 3d shape reconstruction via 2.5 d sketches. Advances in Neural Information Processing Systems **30**, 540–550 (2017)
39. Wu, J., Zhang, C., Xue, T., Freeman, W.T., Tenenbaum, J.B.: Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. pp. 82–90 (2016)
40. Wu, J., Zhang, C., Zhang, X., Zhang, Z., Freeman, W.T., Tenenbaum, J.B.: Learning shape priors for single-view 3d completion and reconstruction. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 646–662 (2018)
41. Xiang, Y., Mottaghi, R., Savarese, S.: Beyond pascal: A benchmark for 3d object detection in the wild. In: IEEE winter conference on applications of computer vision. pp. 75–82. IEEE (2014)
42. Xie, H., Yao, H., Sun, X., Zhou, S., Zhang, S.: Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2690–2698 (2019)
43. Xie, H., Yao, H., Zhang, S., Zhou, S., Sun, W.: Pix2vox++: multi-scale context-aware 3d object reconstruction from single and multiple images. International Journal of Computer Vision **128**(12), 2919–2935 (2020)
44. Xu, Q., Wang, W., Ceylan, D., Mech, R., Neumann, U.: Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. Advances in Neural Information Processing Systems **32**, 492–502 (2019)
45. Yang, S., Xu, M., Xie, H., Perry, S., Xia, J.: Single-view 3d object reconstruction from shape priors in memory. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3152–3161 (2021)
46. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
47. Zhu, L., Yang, Y.: Compound memory networks for few-shot video classification. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 751–766 (2018)